ANÁLISIS MORFOLÓGICO CON HERRAMIENTAS INFORMÁTICAS. RECONOCIMIENTO DE NOMBRES EN TEXTOS DE ESPAÑOL CON EL SISTEMA NOOJ*

Carolina Paola Tramallino
Universidad Nacional de Rosario, Argentina
carotramallino@hotmail.com

Recibido: 11/05/2013 - Aceptado: 12/06/2013

Resumen: Este trabajo tiene como objetivo mostrar los alcances de la lingüística computacional en el uso de herramientas informáticas para el análisis automático morfológico. Se describen dos programas: por un lado, Smorph, *software* creado por Gabriel Bes, cuya formalización refiere al lema y terminaciones; por otro, el sistema Nooj, diseñado por Marx Silverstein para realizar el análisis morfológico, sintáctico y semántico de lenguas naturales. Debido a que este aún no posee datos lingüísticos correspondientes al español, se mostrará la adaptación de los modelos correspondientes a la categoría nombre, declarados en Smorph para la creación de gramáticas y diccionarios en español, necesarios en Nooj.

Palabras clave: Lingüística computacional, análisis automático morfológico, Smorph, Nooj.

MORPHOLOGICAL ANALYZATIONS WITH INFORMATION TOOLS. RECOGNITION OF NAMES IN SPANISH TEXTS USING THE NOOJ SYSTEM

Abstract: The objective of this research work is to show the scope of Computational Linguistics in the use of information tools for the morphological automatic analysis. Two programs are described: on one hand, Smorph, a software created by Gabriel Bes, whose formalization makes reference to headword and endings; on the other hand, Nook system, designed by Marx Silverstein to make the morphological, syntactic and semantic analyses of natural languages. Due to the fact that this system still does not have linguistic data corresponding to Spanish, an adaptation of the models that belong to the noun category, as they are stated in Smorph for the creation of grammars and dictionaries in Spanish will be shown.

Key words: Computational Linguistics, Automatic Morphological Analysis, Smorph, Nooj

---

* Este artículo hace parte de las actividades del grupo de investigación «Infosur: investigación y desarrollo» de la Universidad de Rosario, Argentina.

1. Introducción

El presente trabajo surge como respuesta a un problema metodológico planteado en mi labor como investigadora dentro del campo de estudio correspondiente a la lingüística computacional, que radica en el uso restringido de las herramientas informáticas. Tanto SMORPH[1] como MPS (Módulo Post Smorph), herramientas que utilizo para detectar automáticamente expresiones y estructuras propias de la interlengua de aprendientes de español, son programas de prueba que no están a la venta ni tienen licencia para ser divulgados. Por lo tanto, el objetivo propuesto es aprender el funcionamiento de una nueva herramienta disponible en internet y lista para ser instalada: me refiero al sistema Nooj.

Este programa de libre acceso fue creado en 2002 y es una herramienta para el tratamiento de lenguas naturales desarrollada por Max Silberstein. Actualmente cuenta con un foro, que puede consultarse en la página *http://groups.yahoo.com/group/nooj-info*, en donde los usuarios intercambian opiniones y participan de congresos anuales que son organizados por su creador. El aporte personal reside en poder realizar en el futuro una adaptación de la formalización lingüística que utilizo en este programa con la finalidad de poder ejecutar el análisis, la búsqueda de información y la comprobación de hipótesis en textos de español producidos por aprendientes de segunda lengua. Pero para llevar a cabo este desafío es fundamental la creación de gramáticas y diccionarios propios de la lengua española. Para tal fin recurrimos a la información lingüística declarada en los archivos del programa Smorph, el cual está diseñado para disponer de esa información y adaptarla para el análisis de cualquier lengua. A partir de los modelos creados para sustantivos, adjetivos y verbos se procederá a la adaptación de estos de acuerdo a los requerimientos de algoritmos de Nooj.

Por lo tanto, me referiré en primer lugar a las características y funcionamiento de Nooj, para pasar luego a la descripción de Smorph y explicar de esta forma la adaptación realizada en cuanto a modelos y rasgos morfosintácticos aplicados al español. Por último, mostraré cómo funciona el *software* para el reconocimiento de nombres en un texto literario y cómo realiza el análisis automático, y mencionaré las posibles aplicaciones en la detección automática de estructuras de interlengua. Considero que esta herramienta no solo permitirá

1 Salah Aït-Mokhtar desarrolló en 1995 el programa Smorph (Segmentation et Morphologie) y en 1999 Faiza Abbaci produjo el módulo post-Smorph (MPS).

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

analizar producciones de aprendientes de segundas lenguas sino también ser explotada con un fin didáctico.

2. Sistema Nooj

Nooj se puede descargar libremente del sitio web *http://www.nooj4nlp.net*. El usuario puede emplearlo, entre otras aplicaciones, para:

- Análisis de textos literarios.
- Investigación y extracción desde diarios o corpus técnicos.
- Formalización de fenómenos lingüísticos.
- Aplicaciones computacionales (análisis automático de textos).

Como se especifica en su manual (Silberztein, 2003), Nooj incluye herramientas para crear y mantener fuentes lexicales así como gramáticas sintácticas y morfológicas. Estas maquinarias, al encontrarse integradas, permiten que se desarrollen las operaciones morfológicas sobre los ítems mientras se realiza un análisis sintáctico. Un ejemplo de ello puede ser realizar la transformación de oraciones en frases pasivas o controlar la concordancia morfosintáctica.

Puede procesar textos y corpus de cientos de archivos de texto, y lo más importante es que representa una herramienta multilingüe. Entre las 46 lenguas referidas encontramos: inglés, francés, portugués, alemán, italiano y ruso. En español aún no están cargados los datos lingüísticos correspondientes a gramáticas y diccionarios y por eso, actualmente, nuestro equipo de trabajo está ocupándose de crear los modelos y rasgos correspondientes a las diferentes clases gramaticales, siguiendo las indicaciones de Max Silberstein, adaptando y actualizando la información lingüística declarada en los archivos de texto (.txt) del *software* Smorph. Tal propósito es viable ya que las gramáticas y diccionarios de Nooj son fáciles de crear y de adecuarse. Ambos programas incluyen herramientas que sirven para chequear, limpiar, modificar, conservar y compartir los archivos.

2.1. Tareas de reconocimiento

Los usuarios pueden desarrollar extractores para identificar unidades semánticas en textos extensos, tales como nombres de personas, ubicaciones, fechas, expresiones técnicas o financieras, entre otras.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

Los paradigmas inflexión-derivación son formalizados como bibliotecas de gráficos **(grafos)** estructurados o reglas basadas en textos.

Las gramáticas sintácticas se utilizan para agregar anotaciones al texto o para desambiguar expresiones o estructuras.

Existen dos tipos de recursos lingüísticos:

- *Diccionarios* (archivos con terminación **.dic**) :

Asocian palabras o expresiones a:

- Un conjunto de información, tal como una categoría (por ejemplo, verbo).
- Uno o más paradigmas de inflexiones o derivaciones (por ejemplo, cómo conjugar o nominalizar verbos).
- Una o más propiedades sintácticas.
- Una o más propiedades semánticas (por ejemplo, clases distribucionales como "+Humano").

- *Gramáticas:* se usan para representar una gran cantidad de fenómenos lingüísticos, desde niveles ortográficos y morfológicos a niveles sintagmáticos. Contiene tres tipos de gramáticas:
  - Gramáticas derivacionales o inflexionales (archivos con terminación **.nof**), que se usan para representar las propiedades inflexionales, por ejemplo, conjugaciones o derivaciones como las nominalizaciones de las entradas léxicas. Estos modelos pueden declararse de forma gráfica o mediante reglas.
  - Gramáticas lexicales, ortográficas, morfológicas o terminológicas (archivos con terminación **.nom**). Se usan para representar conjuntos de tipos de palabras y asociarlas con información léxica; por ejemplo: estandarizar la ortografía de palabras o de variantes, reconocer neologismos, asociar expresiones sinónimas. Este operador de disyunción permite comenzar varias búsquedas de extracción de información simultáneamente, tanto para formas flexionales de la misma palabra como también para localizar variaciones ortográficas, nombres y cadenas léxicas como: Rosario | la ciudad; variaciones terminológicas y formas derivadas morfológicamente como, por ejemplo: Argentina | argentino | argentinismo. Además, encuentra expresiones que representan los mismos conceptos como: (crédito | débito | visa) tarjeta + Master.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

- Gramáticas semánticas o sintácticas (archivos con terminación **.nog).** Son utilizadas para reconocer o anotar expresiones en textos como, por ejemplo, etiquetar tanto a frases nominales como a ciertas expresiones sintácticas o idiomáticas. Además, se emplean para extraer o desambiguar palabras filtrando algunos ítems léxicos o anotaciones sintácticas en el texto.

3. Funcionamiento de Smorph

Smorph es un analizador y generador morfosintáctico, que realiza la *tokenización* y el análisis morfológico en una sola etapa, obteniendo como resultado las formas correspondientes a un lema (o a un subconjunto de lemas) con los valores adecuados. Desarrollado en el Gril (Groupe de Recherche dans les Industries de la Langue), bajo la dirección de Gabriel G. Bès, este *software* no solo permite el análisis morfológico y el etiquetado de los textos, sino que también tiene en cuenta el análisis sintáctico. Además, en el caso de palabras desconocidas o no reseñadas en el diccionario, consigue describir la categoría de la palabra a partir de su terminación morfológica. Esto es fundamental a la hora de trabajar con textos de aprendientes del español, en los que se deberá detectar y analizar las formas idiosincrásicas propias de la interlengua.

Cabe destacar que la información utilizada por Smorph está separada de la maquinaria algorítmica, por lo tanto permite adaptarla al uso que necesite darse, de modo tal que con el mismo *software* se puede tratar cualquier lengua e incluso la interlengua de aprendientes de español, si se cambia la información lingüística declarada en sus archivos (Solana, Beltrán y Tramallino, 2009; Tramallino, 2009).

Esta herramienta compila, minimiza y compacta la información lingüística de modo que quede disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos:

- **Códigos Ascii**: en este archivo se especifican, entre otros, los caracteres separadores y las equivalencias entre mayúsculas y minúsculas.
- **Entradas**: funciona como un diccionario lingüístico. Aquí se ingresan los ítems léxicos que deben ir acompañados por un indicador del modelo correspondiente. Por ejemplo: corazón @n12. Este indicador de modelo es la conexión con el correspondiente archivo.

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

- **Terminaciones:** es fundamental declarar todas las terminaciones que son necesarias para definir los modelos de flexión. Dichos caracteres expresan un rasgo o conjunto de rasgos determinados. Por ejemplo, algunas terminaciones para nombres son: o, a, os, as, s, es, z, etc.
- Si en la definición de un modelo se especifica una terminación no declarada en este archivo, el programa emite un mensaje de error. Las terminaciones se declaran una a continuación de otra, separadas por un punto. Es posible declarar una terminación vacía mediante el carácter «@» y una **terminación distinguida** asociando a una terminación la definición morfológica correspondiente.
- En **Rasgos** se organizan jerárquicamente las etiquetas, por ejemplo, nombre, adjetivo, verbos, etc. Asimismo, se puede incorporar la etiqueta que indica, por ejemplo, el tipo de nombre, y se agregan los rasgos de concordancia, género y número.
- **Modelos:** En este archivo se introduce la información correspondiente a los modelos de flexiones morfológicas. Un modelo de flexión agrupa todas las flexiones de una misma clase de palabras. Esto se describe asociando a un conjunto de terminaciones el correspondiente conjunto de definiciones morfológicas. El esquema para definir los modelos es el siguiente:

  <nombre_modelo> -<cantidad de caracteres a sustraer>

  <terminación 1> <definición morfológica para terminación 1>

Estos son introducidos mediante @, que indica el lugar en donde se ubica la raíz a la cual se agregan las terminaciones. Además, se consigna la información morfológica. Por ejemplo, para un modelo de sustantivo como *niño*:

```
@n4      -1
+o       nom/masc/sg
+a       nom/fem/sg
+os      nom/masc/pl
+as      nom/fem/pl
```

Para un sustantivo que se flexiona como la palabra futón, tenemos por ejemplo:

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

```
@n12      -2
+ón       nom/masc/sg
+ones     nom/masc/pl .
```

Aclaraciones: nom (nombre), masc (masculino), fem (femenino), sg (singular), pl (plural).

Se declara en primer lugar el nombre del modelo, indicado con el número, 4 y 12 en los ejemplos. Luego se declara la cantidad de caracteres que hay que sustraer a la forma lematizada. Este valor debe ser una cifra entre 0 y 9 y estar precedida del signo «-». Para formar el femenino de *niño*, por ejemplo, debe sustraerse el último carácter que es la *o*. En el caso de *futón* se quitan dos caracteres para poder eliminar la tilde que no se conserva en el plural. En tercer lugar, se ingresa la terminación, la cual debe estar declarada previamente en el archivo Terminaciones. La información morfológica corresponde a una cadena de caracteres sin espacios en blanco introducida por el signo + que indica adición al lema, por ejemplo: +ón.

4. Implantación en máquina de la categoría nombre en español

Para llevar a cabo esta tarea se utilizarán los modelos morfológicos de Smorph para luego trasladarlos a los archivos de Nooj.

4.1. Modelos Smorph

Para trasladar los modelos morfológicos de flexión correspondientes a los nombres, se extraen del archivo *entradas* todos los lemas que tengan esa asignación de categoría, por ejemplo:

```
implosión       @n13 .
importación     @n13 .
importancia     @n27 .
importe         @n1 .
imposibilidad   @n6 .
impotencia      @n2 .
```


LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

Luego se chequean los números de modelos correspondientes en el archivo del mismo nombre y se le agrega un ejemplo a cada uno de ellos para visualizar el tipo de flexión que sigue. Por ejemplo:

```
@n1       -0 (abrigo)
    +@          nom/masc/sg
    +s          nom/masc/pl .
@n2       -0 (mesa)
    +@          nom/fem/sg
    +s          nom/fem/pl .
@n3       -0 (accionista)
    +@          nom/_/sg
    +s          nom/_/pl
@n4       -1 (perro)
    +o          nom/masc/sg
    +a          nom/fem/sg
    +os         nom/masc/pl
    +as         nom/fem/pl
```

4.2. Modelos Nooj

Nooj describe la flexión mediante comandos, por ejemplo, <E> indica cadena vacía, es decir, que el lema quedará exactamente igual; para quitar caracteres, empleamos el operador <B>, que borrará el último carácter, y su variante <B2> para realizar la operación dos veces. Para extraerle el acento a la letra actual utilizamos <A>.

Estos comandos facilitan la tarea de creación de modelos tanto para sustantivos como para adjetivos, ya que para declarar un ítem léxico se debe deducir el modelo de flexión que sigue y emplear los comandos ya dispuestos en el programa para flexionar esa palabra según las variaciones que presente en cuanto a género y número. Por lo tanto, al querer trasladar los modelos asignados en Smorph para los nombres del español (producto del trabajo del equipo INFOSUR) de acuerdo a los requerimientos de la herramienta Nooj, notamos que podemos agrupar varios de los modelos de aquél en uno solo de los que necesita éste para reconocer y realizar el análisis automático. Por

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

ejemplo, a lexemas como *montón, diván, retén, budín, atún, anís, compás, interés,* que poseen género masculino y realizan el plural agregando "es" sin tilde, les corresponden los siguientes modelos de Smorph:

@n12 -2 (montón)
+ón nom/masc/sg
+ones nom/masc/pl .
@n16 -2 (diván)
+án nom/masc/sg
+anes nom/masc/pl .
@n17 -2 (retén)
+én nom/masc/sg
+enes nom/masc/pl .
@n18 -2 (budín)
+ín nom/masc/sg
+ines nom/masc/pl .
@n20 -2 (betún)
+ún nom/masc/sg
+unes nom/masc/pl .
@n21 -2 (anís)
+ís nom/masc/sg
+ises nom/masc/pl .
@n22 -2 (obús)
+ús nom/masc/sg
+uses nom/masc/pl .
@n40 -2 (compás)
+ás nom/masc/sg
+ases nom/masc/pl .
@n46 -2 (interés)
+és nom/masc/sg
+eses nom/masc/pl.

En Nooj, en cambio, todos esos modelos pueden reducirse a uno utilizando los siguientes comandos **=<E>/masc+sg | <A> es/masc+pl,** que significan que para formar el masculino singular la palabra queda exactamente igual; el

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

segundo comando: <A> indica que se le quite el acento al último carácter y que luego se agregue «es» para formar el masculino singular. Este tipo de nombre no posee formas en femenino.

Una de las razones que posibilitan la utilización de una cantidad menor de modelos es el hecho de que para extraer la tilde en Smorph debe quitarse el carácter que lleva la tilde y todos los que le siguen hasta el final de la palabra y, luego, agregar el mismo segmento con la vocal sin tilde; en cambio, en Nooj existe la posibilidad de eliminar únicamente la tilde sin cambiar el grafema que lo porta.

Para llenar el diccionario de Nooj bastará con extraer los nombres cargados en las entradas de Smorph y cambiarles la designación de modelo por la forma que corresponda según los modelos de flexión creados a partir de los comandos existentes en Nooj.

5. Reconocimiento de nombres en un texto literario

Para ofrecer una pequeña muestra del trabajo que realiza Nooj procuraré que reconozca los nombres en un texto literario y, además, que analice morfológicamente enunciados.

Para ambas tareas es necesario crear nuevos textos e ingresar los ítems léxicos, modelos de flexión y categorías con las que debe operar el *software*. Estos se consignan en tres archivos de diferente extensión: los modelos se ubican en el archivo Gramática («Grammar») que posee extensión .nog, los lemas se ingresan en el archivo Diccionario («Dictionary») con extensión .dic y, por último, las categorías gramaticales con sus rasgos, por ejemplo, Nombre: género/ número, en el archivo Propiedades («Properties ´ definition»), que tiene extensión .def

5.1. Propiedades

En este archivo deben consignarse en mayúscula los rasgos que se utilizan para etiquetar las entradas del diccionario con sus respectivos valores que se declaran en minúscula. Estos pueden trasladarse desde el archivo Rasgos de Smorph pero deben cargarse de la siguiente manera:

CATEGORÍAS = AJ + V +N + AV + PREP +DET
N_género = masc + fem;

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

N_número = sg + pl + inv;
AJ_dialecto = RIOP + LUNF;

Como puede verse, ingresamos las siguientes etiquetas.: adjetivo (AJ), verbo (V), nombre (N), adverbio (AV), preposición (PREP), determinante (DET).

5.2. Gramática

En Gramática deben declararse los siguientes siete modelos, que abarcan a los más de cien nombres que aparecen en el texto a analizar:

FELICIDAD = <E>/fem+sg | es/fem+pl;
TERROR = <E>/masc+sg | es/masc+pl;
MESA = <E>/fem+sg | s/fem+pl;
MUCHACHO = <E>/masc | <B> a/fem <E>/sg | s/pl;
INFIERNO = <E>/sg | s/pl;
SECCIÓN = <E>/fem+sg | <A> es/fem+pl;
VEZ = <E>/fem+sg | <V> es/fem+pl;

5.3. Diccionario

Para cada idioma, Nooj tiene acceso a un diccionario en el que cada palabra de esa lengua es una entrada y se asocia a cierta información morfológica, por lo general a su inflexión o paradigmas derivativos. El paradigma flexivo dice a Nooj cómo flexiona la entrada léxica, es decir, cuáles son sus formas conjugadas (si se trata de un verbo) y su femenino y plural (para los nombres en las lenguas románicas). Para ingresar los nombres del texto elegido debemos seleccionar sustantivos que funcionen como referencia de los siete modelos flexivos declarados en el archivo Gramática, de esta forma:

felicidad,N+FLX=FELICIDAD
mesa,N+FLX=MESA
muchacho,N+FLX=MUCHACHO
terror,N+FLX=TERROR
infierno,N+FLX=INFIERNO
sección,N+FLX=SECCIÓN
vez,N+FLX=VEZ

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

Y a continuación ingresamos todos los ítems indicando el modelo de flexión que siguen, de la siguiente manera:

(Se muestra un fragmento de la totalidad de los ítems ingresados)

amor,N+FLX=TERROR	tarde,N+FLX=MESA
penuria,N+FLX=MESA	costumbre,N+FLX=MESA
cita,N+FLX=MESA	alma,N+FLX=MESA
tierra,N+FLX=MESA	novela,N+FLX=MESA
puerta,N+FLX=MESA	entrega,N+FLX=MESA
presencia,N+FLX=MESA	hermano,N+FLX=MUCHACHO
balanza,N+FLX=MESA	biógrafo,N+FLX=MUCHACHO
culpa,N+FLX=MESA	novio,N+FLX=MUCHACHO
trampa,N+FLX=MESA	pecho,N+FLX=INFIERNO
medida,N+FLX=MESA	tango,N+FLX=INFIERNO
miseria,N+FLX=MESA	día,N+FLX=INFIERNO
cuna,N+FLX=MUJERCITA	vestido,N+FLX=INFIERNO
minuto,N+FLX=INFIERNO	paquete,N+FLX=INFIERNO
espanto,N+FLX=INFIERNO	tendón,N+FLX=SECCIÓN
pibe,N+FLX= MUCHACHO	arroz,N+FLX=VEZ
asunto,N+FLX=INFIERNO	voz,N+FLX=VEZ
rostro,N+FLX=INFIERNO	madurez, N+FLX=VEZ
asombro,N+FLX=INFIERNO	escasez, N+FLX=vez
organización,N+FLX=SECCIÓN	preocupación,N+FLX=SECCIÓN
inflexión,N+FLX=SECCIÓN	indignación,N+FLX=SECCIÓN
razón,N+FLX=SECCIÓN	privación,N+FLX=SECCIÓN

5.4. Resultados

Para localizar una expresión en un determinado texto, debe abrirse este, que habrá sido guardado dentro de la carpeta «sp» correspondiente al español a partir de los siguientes pasos: Open / Text, se selecciona y se pincha en *abrir*, luego en la ventana TEXT se elige Linguistic Analysis y luego Locate, aquí se coloca la expresión o expresiones en mayúscula y entre <>, por ejemplo

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

<BRAZO>, eligiendo entre distintas opciones de búsqueda: solo en cien oraciones que servirán para extraer datos porcentuales o en todas las oraciones.

5.4.1. Salida de Nooj

(...): la costumbre de llevar el atado siempre del brazo opuesto:
Cuando estas muchachas cumplieron ocho o nueve años,
tuvieron que cargar un hermanito en los brazos.
Usted, como yo, debe haber visto en el arrabal estas mocosas que cargan un pebetito
en el brazo y que se
pasean por la vereda rabiando contra el mocoso, y vigiladas por la madre que salpicaba agua en la batea.

6. Análisis morfológico

Para el análisis morfológico debemos ir a New / Text para crear un nuevo texto y seleccionar la lengua, en este caso «sp» que corresponde a español (*spanish*), y escribir en el cuadro de texto abierto la expresión, enunciado o texto que queramos analizar, por ejemplo: *razón* (que sigue el modelo de *sección*) y luego ir a TEXT y seleccionar Linguistic Analysis y por último pinchar en el cuadro Show Text Annotation Structure.

6.1. Salida de Nooj

+s **razón**
+p **razones**

Para un sustantivo como *novio*, que sigue el modelo de flexión de *muchacho*, la salida de Nooj arroja:

m+s **novio**
f+s **novia**
m+p **novios**
f+p **novias**

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

7. Posibles aplicaciones de Nooj en el campo de la adquisición de segundas lenguas

Con respecto a la primera etapa del análisis de la interlengua que tiene que ver con la normalización ortográfica, Nooj puede detectar repeticiones de letras. A diferencia de Smorph, que no reconoce expresiones que tengan alguna alteración ortográfica y las analiza como palabras desconocidas, Smorph solo reconoce palabras que están incluidas en sus diccionarios. Por lo tanto, para que reconozca formas que presentan alguna alteración ortográfica es necesario agregarlas al diccionario; en cambio, Nooj permite desarrollar gramáticas que reconozcan determinadas alteraciones sistemáticas, por ejemplo, las palabras escritas con doble consonante como: «possible» (inglés), «*interessante*» (alemán), «*difficil*» (italiano), «*apprender*» o «*dessarrollo*» (portugués), usuales en producciones de los aprendientes. Estas pueden ser tratadas mediante la siguiente gramática: Nooj, indicando qué palabras presentan repeticiones, como se explica en el Manual: "Todas las palabras reconocidas serán anotadas con la categoría de "REP".

Otra función de utilidad de la herramienta es que no solo localiza la forma solicitada (en términos computacionales, diríamos la concordancia de una secuencia que representa la totalidad de sus apariciones en el contexto), sino que también indica la ubicación espacial de la expresión brindando el nombre del texto o el capítulo, si se trata de un libro, en el que se encuentra. Esto facilita la tarea a la hora de trabajar con grandes corpus pertenecientes a distintos sujetos que a la vez poseen distintas lenguas de origen:

> Las concordancias de Nooj se muestran en cuatro columnas: cada ocurrencia se presenta en la columna del medio, en su contexto adecuado. Si lo que se indexa es un corpus (es decir, un conjunto de archivos de texto), la primera columna muestra el nombre del archivo de texto en el que cada ocurrencia se produce. Se puede variar el tamaño del contexto izquierdo y derecho, así como el orden de la concordancia (Silberztein, 2003: 123).

Para detectar estructuras sintácticas en donde se produce una asignación incorrecta de género en el sustantivo y, por lo tanto, falla la concordancia entre determinante y sustantivo o entre este y el adjetivo (Méndez, 2009), debe recurrirse al Módulo Post Smorph (MPS), que trabaja con reglas de reagrupamiento, y crear una regla que establezca, por ejemplo, que un determinante masculino

LINGÜÍSTICA Y LITERATURA
ISSN 0120-5587
No. 64, 2013, 33-48

más un sustantivo femenino es una estructura propia de interlengua. Con Nooj es más sencillo porque:

- El sistema morfológico de Nooj tiene otros dos operadores que pueden ser utilizados para comprobar la igualdad (o la desigualdad) de dos propiedades: el operador de igualdad "=" y el operador de desigualdad "!=". Por ejemplo, la restricción de concordancia: <NomNumber!=AdjNumber> comprueba que no haya concordancia.
- En los diccionarios de Nooj todas las entradas están asociadas a una categoría morfosintáctica, por lo tanto, en las expresiones regulares pueden utilizarse categorías. Por ejemplo, para localizar todas las secuencias que contienen cualquier forma asociada con el lema «be», seguido por una preposición y, a continuación, un nombre, debemos indicar que busque la siguiente expresión: <be> <PREP> <N>. Para trabajar, por ejemplo, los verbos livianos que aparecen en construcciones de interlengua, podríamos indicarle que localice <dar> <DETER> <N> y entonces buscará expresiones como: *dar un beso*, *dio una ducha,* etc.
- Finalmente, es pertinente aclarar que con Nooj los lingüistas pueden crear libremente sus propias categorías y códigos acorde al trabajo de investigación que estén llevando a cabo. En mi caso tengo la posibilidad de añadir, por ejemplo: VEINTERL, para identificar a un verbo propio de la interlengua de aprendientes de español.

8. Consideraciones finales

En este trabajo se presentó la herramienta Nooj, describiendo su funcionamiento y particularidades. Luego se explicó el trabajo realizado a partir de Smorph para lograr el análisis automático y el pasaje de modelos morfológicos, propiedades de las categorías y entradas de ítems léxicos de un programa a otro según los requerimientos algorítmicos de ambos. A continuación se mostró la implantación en máquina de los nombres aparecidos en el texto *La muchacha del atado*, de Roberto Arlt, y el análisis de los mismos para hacer visible la tarea realizada por Nooj. La conclusión es que, además de ser un programa informático que brinda disponibilidad y gratuidad para su empleo, ya que puede descargarse *online*, cuenta con comandos que facilitan la creación de una cantidad reducida de modelos morfosintácticos que abarquen la totalidad

de nombres y adjetivos del español. Lo más interesante es que, además, abre un abanico de posibilidades más amplio para el tratamiento automático de la interlengua, y esto gracias a que pueden modificarse y adaptarse sus archivos, pueden generarse nuevas categorías y a que permite no solo la detección de estructuras sintácticas concordantes y no concordantes sino también el análisis desde categorías que refieren al campo léxico del lenguaje. El objetivo perseguido es continuar estudiando este *software* y completar la gramática del español en lo que refiere a verbos (Bonino, 2011) y otras categorías para luego poder adaptarla para el análisis de la interlengua de aprendientes de español, con la expectativa de que en un futuro pueda servir como herramienta didáctica para que docentes y estudiantes puedan chequear las producciones y corregirlas según el objetivo educativo que se pretenda alcanzar.

Referencias bibliográficas

1. Aït-Mokhtar, S. & Rodrigo Mateos, J. L. (1995). Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. *SEPLN*, 17, 29-41.
2. Bonino, R. (2011). Una propuesta para la implantación de la morfología verbal del español. *INFOSUR*, 5, 79-86.
3. Méndez, B. (2009). Análisis automático de la interlengua: asignación de género y número diferentes a la lengua estándar en el sintagma nominal núcleo. En *La interlengua de aprendientes de español como L2*. Centro de Estudios de Adquisición del Lenguaje. Rosario: UNR.
4. Silberztein, M. (2003). *Nooj Manual*. Traducción al español a cargo de Rodolfo Bonino. Recuperado de: <http://www.nooj4nlp.net/NooJManual.pdf> en 11/2012.
5. Solana, Z., Beltrán, C., & Tramallino, C. (2009). La implantación en máquina de la interlengua de los aprendientes de español como L2: los sufijos formadores de nombres. En *La interlengua de aprendientes de español como L2: aportes de la Lingüística informática*, 23-28. Rosario: Centro de Estudios de Adquisición del Lenguaje UNR-Ed. Juglaría.
6. Tramallino, C. (2009). Formas verbales irregulares en la interlengua de aprendientes de español como L2. En *La interlengua de aprendientes de español como L2: aportes de la Lingüística informática*, 42-44. Rosario: UNR. Centro de Estudios de Adquisición del Lenguaje-Ed. Juglaría.